Vol.2021-MUS-132 No.24
2021/9/17

非負値行列因子分解を用いた被り音の抑圧

溝渕 悠朔[1]　北村 大地[1]　中村 友彦[2]　猿渡 洋[2]　高橋 祐[3]　近藤 多伸[3]

概要：音楽の生演奏を録音する場合，各音源に近接するようにマイクロホンを配置することが一般的である．これは，各音源のみの音響信号を得ることを目的としているが，実際には他の音源からの音も少なからず混入してしまう．これは一般に被り音と呼ばれる．本研究では，音楽信号の録音時に生じる被り音の抑圧を目的として，非負値行列因子分解を用いた新しい手法を提案する．本手法では，周波数毎の被り音成分の漏れゲインに対してガンマ事前分布を導入し，混合行列を教師なし推定する．推定された混合行列から Wiener フィルタを構成することで，被り音を抑圧した信号を得ることができる．シミュレーションによる音楽信号の被り音抑圧実験では，提案手法が従来手法と比べて高い精度で被り音を抑圧できることを示す．

1. はじめに

音楽の生演奏を録音する場合，音源数以上のマイクロホンを配置することが一般的である．特に，Fig. 1 に示すように，楽器本体やボーカル，アンプ等の各音源に近接させてマイクロホンを配置することが多い．これらのマイクロホンは，近接している音源（以後，目的音源と呼ぶ）からの音響信号のみを観測することを意図して配置されている．しかし，通常はこれらの近接マイクロホンに，目的音源以外の音源（以後，非目的音源と呼ぶ）の音が漏れて混入してしまう．この問題は，一般にクロストークや被り音等と呼ばれている．

ライブ演奏のステージ上でのミキシングでは，サウンドエンジニアが各音源の音量バランスや音質を調節し，会場用スピーカ及びモニタースピーカを通して，それぞれ観客及び演奏者に提供する．このようなサウンドエンジニアによる音量及び音質の調整は sound reinforcement（SR）と呼ばれる．近接マイクロホンから得られる各観測信号に被り音が含まれている場合は SR が困難となり，会場の音質や演奏の質を低下をさせてしまう可能性がある．また，録音時の被り音は，ライブ演奏後のリミックスにおいても悪影響を及ぼす問題がある．以上の理由から，サウンドエンジニアは録音時の被り音を極力回避するために，近接マイクロホンの配置を慎重に行う必要がある．音楽スタジオでの収録では，音源間に遮音壁を設けることや，吸音材等で

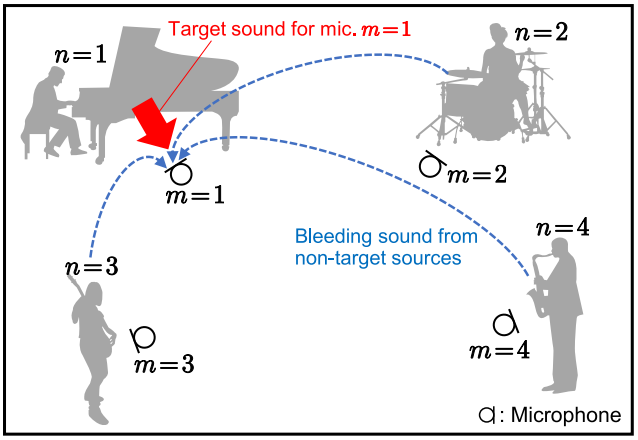

Fig. 1 Spatial arrangement of sources and close microphones, where $M = N = 4$. Target sound is contaminated with bleeding sound from other non-target sources.

音の反射を抑えることで，被り音の低減を試みることもあるが，同じ環境で複数の音源が同時に鳴る状況では，被り音を完全に防ぐことはほとんど不可能である．

複数のマイクロホンで観測された信号に対する被り音の抑圧は，多チャネル音源分離（multichannel audio source separation: MASS）[1], [2], [3], [4] と呼ばれる問題と類似している．但し，次に示す点において MASS とは異なる特徴がある．

(a) 観測信号の信号対ノイズ（signal-to-noise: SN）比は比較的高い．ここで，SN 比における信号は目的音源の成分を指し，ノイズは非目的音源からの混入成分である．

(b) 各マイクロホンにおける目的音源は既知である（観測信号の各チャネルには音源のラベルが付与される）．

[1] 香川高等専門学校
National Institute of Technology, Kagawa College
[2] 東京大学
The University of Tokyo
[3] ヤマハ株式会社
Yamaha Corporation

(c) 一般的なマイクロホンアレイと異なり，マイクロホン間隔が空間的に離れており（例えば 2 m 以上），観測信号には空間エイリアシングが発生する．
(d) 音楽信号を対象するため，その芸術的価値を低下させないよう高品質な被り音抑圧が要求される．

条件 (a) と (b) は，MASS と比較した際に被り音抑圧の方が容易と考えられる性質である．一方，条件 (c) と (d) は被り音抑圧における困難な点であり，特に条件 (c) の空間エイリアシングはより深刻である．ビームフォーミング [1], [2] やブラインド音源分離（blind source separation: BSS）[5], [6], [7], [8], [9], [10], [11] 等の多くの高品質な MASS はマイクロホン間の位相差を利用しており，空間エイリアシングが生じる場合は音源を抑圧・分離することが極めて困難になる．

観測信号の位相を無視し振幅やパワーのみを用いる MASS もいくつか提案されている [12], [13], [14], [15]. Togami ら [12] は，多チャネル観測信号の周波数毎の時間チャネル領域に対して非負値行列因子分解（nonnegative matrix factorization: NMF）[17], [18] を適用する時間チャネル NMF（time-channel NMF: TCNMF）を提案している．TCNMF では，周波数ビン毎の非負混合行列と各音源のアクティベーションを推定しており，条件 (c) や非同期録音多チャネル信号に対しても音源分離できることが確認されている [13]．また，TCNMF と同じく位相情報を無視した BSS として，線形分離領域多チャネル NMF（linear demixed domain multichannel NMF: DMNMF）[15] がある．DMNMF も TCNMF と同じく周波数ビン毎の非負混合行列を推定する BSS であるが，いずれの手法も音楽信号の被り音抑圧に対する有効性は検討されていない．音楽信号の被り音抑圧を目的とした手法では，周波数毎の非負混合行列の対角成分（近接マイクロホン毎の目的音源のゲイン）と非対角成分（近接マイクロホン毎の非目的音源の漏れゲイン）を事前計測する教師あり手法が提案されている [16]．しかし，録音現場での SR のコストを極力抑えるためには，このような事前計測はあまり望ましくない．また，事前計測された混合行列と実際の状態に大きなミスマッチがある場合，被り音抑圧の性能は著しく低下してしまう．

本研究では，音源やマイクロホンの空間的な配置が不明（ブラインド）の条件下で被り音を抑圧することを目的とする．但し，訓練データとテストデータのミスマッチを避けるために，ソロ演奏のデータセット等の学習データや混合系の事前計測等は用いない．従って，教師あり学習に基づいた手法 [19], [20], [21], [22], [23] は本研究の対象外である．本稿では，位相に依存しないブラインド被り音抑圧手法として，TCNMF の改良手法を新たに提案する．被り音の相対的な漏れゲインをモデル化するために，新たに周波数毎の非負混合行列の対角及び非対角成分に事前分布を導入する．この事前分布の導入については，Cemgil [24] によって提案された最大事後確率（maximum a posteriori: MAP）推定 NMF に基づいている．

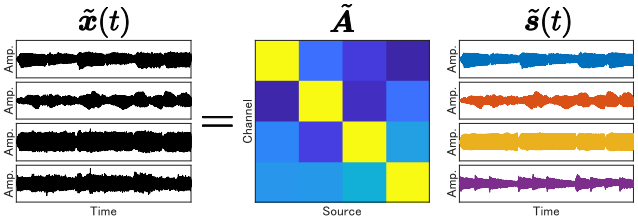

Fig. 2 Instantaneous mixture model for bleeding-sound reduction, where $M = N = 4$. Color brightness in mixing matrix $\tilde{\boldsymbol{A}}$ shows amplitude level of each element (brighter is larger). Due to close miking setup, diagonal elements in $\tilde{\boldsymbol{A}}$ have larger amplitudes compared with off-diagonal elements.

2. 従来手法

2.1 混合モデル

M 及び N をそれぞれマイクロホン数及び音源数とし，音源信号，観測信号，及び推定信号をそれぞれ次式で表す．

$$\tilde{\boldsymbol{s}}(t) = [\tilde{s}_1(t), \cdots, \tilde{s}_n(t), \cdots, \tilde{s}_N(t)]^{\mathrm{T}} \in \mathbb{R}^N \quad (1)$$
$$\tilde{\boldsymbol{x}}(t) = [\tilde{x}_1(t), \cdots, \tilde{x}_m(t), \cdots, \tilde{x}_M(t)]^{\mathrm{T}} \in \mathbb{R}^M \quad (2)$$
$$\tilde{\boldsymbol{y}}(t) = [\tilde{y}_1(t), \cdots, \tilde{y}_n(t), \cdots, \tilde{y}_N(t)]^{\mathrm{T}} \in \mathbb{R}^N \quad (3)$$

ここで，$t = 1, 2, \cdots, T$，$n = 1, 2, \cdots, N$，及び $m = 1, 2, \cdots, M$ はそれぞれ離散時間，音源，及びマイクロホンのインデクスである．被り音抑圧では，1 章で述べた録音条件より，混合系は優決定系（$M \geq N$）となる．本稿では，以後 $M = N$ の条件のみ取り扱う．

複数の音源信号が瞬時混合する場合，観測信号及び推定信号はそれぞれ次式でモデル化される．

$$\tilde{\boldsymbol{x}}(t) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{s}}(t) \quad (4)$$
$$\tilde{\boldsymbol{y}}(t) = \tilde{\boldsymbol{W}}\tilde{\boldsymbol{x}}(t) \quad (5)$$

ここで，$\tilde{\boldsymbol{A}} \in \mathbb{R}^{M \times N}$ 及び $\tilde{\boldsymbol{W}} \in \mathbb{R}^{N \times M}$ は，それぞれ時不変な混合行列及び分離行列である．この時不変瞬時混合モデル (4) を Fig. 2 に示す．観測信号 $\tilde{\boldsymbol{x}}(t)$ は，1 章の条件 (b) より各チャネルに目的音源のラベルが付与されていることから，$\tilde{x}_m(t)$ は Figs. 1 及び 2 に示すように，m 番目の音源 $\tilde{s}_m(t)$（$n = m$）の近接マイクロホンの観測信号と定義する．従って，$\tilde{x}_m(t)$ には目的音源 $\tilde{s}_m(t)$ の成分が主に含まれており，同時に非目的音源 $\tilde{s}_{m'}(t)$ の被り音成分が含まれている．ここで，$m' \neq m$ である．このため，$\tilde{\boldsymbol{A}}$ の対角成分の絶対値（目的音源のゲイン）は大きく，非対角成分の絶対値（非目的音源の漏れゲイン）は小さくなる（1 章の条件 (a) に対応）．被り音抑圧は，混合行列 $\tilde{\boldsymbol{A}}$ 又は $\tilde{\boldsymbol{W}} = \tilde{\boldsymbol{A}}^{-1}$ なる分離行列 $\tilde{\boldsymbol{W}}$ を，観測信号 $\{\tilde{\boldsymbol{x}}(t)\}_{t=1}^{T}$ のみ

から推定する問題となる．

実際の録音では，音が各マイクロホンに到達するまでの時間差や録音環境の残響により，複数音源の混合は時不変畳み込み混合となる．本稿では，時不変畳み込み混合をシンプルにモデル化するため，各マイクロホンと各音源間のインパルス応答長（即ち残響長）が，短時間フーリエ変換（short-time Fourier transform: STFT）で用いる窓長よりも短いと仮定する．この仮定により，複数の音源が時不変畳み込み混合された観測信号は，時間周波数領域では周波数毎の時不変複素瞬時混合として表すことができ，観測信号と推定信号はそれぞれ次式となる[*1]．

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij} \tag{6}$$
$$\mathbf{y}_{ij} = \mathbf{W}_i \mathbf{x}_{ij} \tag{7}$$

ここで，$\mathbf{s}_{ij}$，$\mathbf{x}_{ij}$，及び $\mathbf{y}_{ij}$ はそれぞれ次式で定義される．

$$\mathbf{s}_{ij} = [\mathrm{s}_{ij1}, \cdots, \mathrm{s}_{ijn}, \cdots, \mathrm{s}_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N \tag{8}$$
$$\mathbf{x}_{ij} = [\mathrm{x}_{ij1}, \cdots, \mathrm{x}_{ijm}, \cdots, \mathrm{x}_{ijM}]^{\mathrm{T}} \in \mathbb{C}^M \tag{9}$$
$$\mathbf{y}_{ij} = [\mathrm{y}_{ij1}, \cdots, \mathrm{y}_{ijn}, \cdots, \mathrm{y}_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N \tag{10}$$

また，$i = 1, 2, \cdots, I$ 及び $j = 1, 2, \cdots, J$ はそれぞれ周波数ビン及び時間フレームのインデクスであり，$\mathbf{A}_i \in \mathbb{C}^{M \times N}$ は周波数毎の複素混合行列である．さらに，s_{ijn}，x_{ijm}，及び y_{ijn} はそれぞれ，複素数の音源，観測，及び推定スペクトログラム $\mathbf{S}_n \in \mathbb{C}^{I \times J}$，$\mathbf{X}_m \in \mathbb{C}^{I \times J}$，及び $\mathbf{Y}_n \in \mathbb{C}^{I \times J}$ の時間周波数要素である．

一般的なビームフォーミング [1], [2] や BSS [5], [6], [7], [8], [9], [10], [11] は，各マイクロホンへ間の位相差を用いて複素分離行列 $\mathbf{W}_i$ を推定する．マイクロホン間隔が大きく離れている場合は空間エイリアシングが生じるため，ビームフォーミングや BSS では $\mathbf{W}_i$ を正確に推定できなくなる．従って，位相差に基づくビームフォーミングや BSS で被り音を抑圧することは困難となる．

2.2 DMNMF

空間エイリアシングに対処するために，位相を無視しパワーのみ用いる BSS の DMNMF が提案されている [15]. DMNMF は，独立低ランク行列分析（independent low-rank matrix analysis: ILRMA）[4], [10], [11] の位相非依存な手法である．DMNMF では，観測信号を次式でモデル化する．

$$\boldsymbol{x}_{ij}^{.2} \approx \boldsymbol{A}_i \boldsymbol{s}_{ij}^{.2} \quad \forall i, j \tag{11}$$
$$\boldsymbol{A}_i = \mathrm{abs}(\mathbf{A}_i) \in \mathbb{R}_{\geq 0}^{M \times N} \tag{12}$$
$$\boldsymbol{x}_{ij} = \mathrm{abs}(\mathbf{x}_{ij}) \in \mathbb{R}_{\geq 0}^{M} \tag{13}$$
$$\boldsymbol{s}_{ij} = \mathrm{abs}(\mathbf{s}_{ij}) \in \mathbb{R}_{\geq 0}^{N} \tag{14}$$

ここで，ベクトルや行列のドット付き指数演算 $\cdot^{.q}$ 及び絶対値演算 $\mathrm{abs}(\cdot)$ はそれぞれ要素毎の q 乗及び要素毎の絶対値を表す．従って，$\boldsymbol{x}_{ij}^{.2}$ 及び $\boldsymbol{s}_{ij}^{.2}$ はそれぞれ $\{\mathbf{X}_m\}_{m=1}^M$ と $\{\mathbf{S}_n\}_{n=1}^N$ のパワースペクトログラムの時間周波数要素となる．DMNMF では，位相情報を無視したパワースペクトログラム領域において，式 (6) を周波数毎の非負混合行列 $\boldsymbol{A}_i$ で式 (11) として近似する．さらに，各音源のパワースペクトログラムは，NMF を用いて低ランク行列でモデル化される．$\boldsymbol{x}_{ij}^{.2}$ から $\boldsymbol{A}_i$ 及び $\boldsymbol{s}_{ij}^{.2}$ を推定した後に，Wiener フィルタを用いて推定信号 $\mathbf{y}_{ij}$ を復元する．

[*1] 本稿では，ローマン体の信号は複素数を表し，イタリック体の信号は実数を表す．

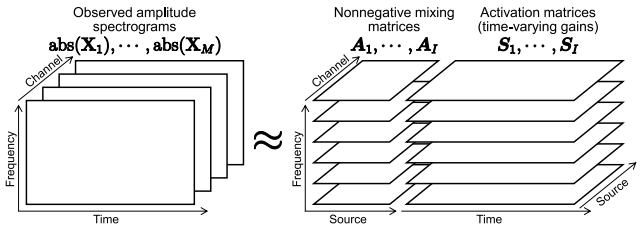

Fig. 3 Decomposition model of TCNMF, where $M = N = 4$ and $I = 6$. Note that $\mathrm{abs}(\mathbf{X}_m)$ is channel-wise time-frequency matrix, but $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ are frequency-wise source-channel and time-source matrices, respectively.

2.3 TCNMF

TCNMF [12] は位相を無視し振幅のみ用いる BSS であり，音声強調に適用されている．一般的な NMF が時間周波数行列を低ランク分解するのに対し，TCNMF は周波数毎の時間チャネル行列を次のように分解する．

$$\boldsymbol{X}_i \approx \boldsymbol{A}_i \boldsymbol{S}_i \quad \forall i \tag{15}$$
$$\boldsymbol{X}_i = [\boldsymbol{x}_{i1} \ \cdots \ \boldsymbol{x}_{ij} \ \cdots \ \boldsymbol{x}_{iJ}] \in \mathbb{R}_{\geq 0}^{M \times J} \tag{16}$$

Fig. 3 に示すように，$\boldsymbol{X}_i$ は振幅領域の周波数ビン毎の時間チャネル観測信号，$\boldsymbol{S}_i \in \mathbb{R}_{\geq 0}^{N \times J}$ は各音源のゲインの時間変化を行ベクトルとして含む行列である．$\boldsymbol{X}_i$ から $\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ を推定した後に，Wiener フィルタを用いて推定信号を復元する．

変数 $\boldsymbol{A}_i$ 及び $\boldsymbol{S}_i$ の推定は，次の最適化問題となる [18].

$$\min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i)$$
$$\text{s.t. } a_{imn}, s_{inj} \geq 0 \quad \forall i, m, n, j \tag{17}$$

ここで，$\mathcal{D}_{\mathrm{KL}}(\cdot|\cdot)$ は次式のように定義される．

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) = \sum_{m,j} \left(x_{imj} \log \frac{x_{imj}}{\sum_n a_{imn} s_{inj}} - x_{imj} + \sum_n a_{imn} s_{inj} \right) \tag{18}$$

式 (18) は，$\boldsymbol{X}_i$ 及び $\boldsymbol{A}_i \boldsymbol{S}_i$ の類似性を測る一般化 Kullback–Leibler（KL）ダイバージェンスである．また，$\mathcal{A}$ 及び $\mathcal{S}$ は，それぞれ $\{\boldsymbol{A}_i\}_{i=1}^I$ 及び $\{\boldsymbol{S}_i\}_{i=1}^I$ の集合であり，x_{imj}，

a_{imn}，及び s_{inj} は，それぞれ $\boldsymbol{X}_i$，$\boldsymbol{A}_i$，及び $\boldsymbol{S}_i$ の要素である．決定系では $\boldsymbol{A}_i$ が $M \times N$ の正方行列であるため，最小化問題である式 (17) は，全ての i に対して $\boldsymbol{A}_i = \boldsymbol{I}$（$\boldsymbol{I}$ は単位行列）という自明解を持つ．この自明解を回避するために，$L_{0.5}$ ノルムに基づくスパース正則化が各時間フレームに対して導入されている [12]．

$$\min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i) + \mu \sum_{i,j} \|\boldsymbol{s}_{ij}\|_{0.5}$$
$$\text{s.t. } a_{imn}, s_{inj} \geq 0 \quad \forall i,m,n,j \tag{19}$$

ここで，μ は正則化のための重み係数，$\boldsymbol{s}_{ij}$ は $\boldsymbol{S}_i$ の時間軸方向のベクトル（$\boldsymbol{S}_i = [\boldsymbol{s}_{i1} \ \cdots \ \boldsymbol{s}_{ij} \ \cdots \ \boldsymbol{s}_{iJ}]$）である．

3. 提案手法

3.1 動機

前述の通り，被り音抑圧では観測信号の位相情報が信用できため，位相に依存しない DMNMF は合理的なアプローチと考えられる．しかしながら，実験で確認する通り DMNMF の最適化は不安定であり，非負混合行列 $\boldsymbol{A}_i$ の高精度な推定は難しい問題である．実際に，文献 [15] では，ステアリングベクトル（$\mathbf{A}_i$ の列ベクトル）の事前情報や位相依存の BSS による事前推定を用いて推定精度を安定させている．一方で TCNMF は，位相情報を用いずにある程度の精度で音声強調ができる [13] が，音楽の BSS や被り音抑圧に対する性能は調査されていない．特に，式 (19) のスパース正則化項 $\sum_{i,j} \|\boldsymbol{s}_{ij}\|_{0.5}$ は，時間周波数領域における W-disjoint-orthogonality [25]（1 つの時間周波数グリッドに生起する音源数は混合信号であっても高々 1 つ）という仮定に基づいている．これは音声の混合信号には妥当だが，時間周波数領域で多分に重なり合う音楽信号では不適切と考えられる．従って，従来の TCNMF で用いられる $\boldsymbol{S}_i$ のスパース正則化は，音楽信号の場合には音質を低下させる可能性がある．

提案手法では，TCNMF における $\boldsymbol{A}_i$ の自明解を避けるために，$\boldsymbol{S}_i$ ではなく非負混合行列 $\boldsymbol{A}_i$ の対角成分と非対角成分をそれぞれ正則化する．本手法は，被り音の相対的な漏れゲインがガンマ事前分布から生成されると仮定した MAP 推定と解釈できる．

3.2 KL ダイバージェンスに基づく NMF の生成モデル

Cemgil は，KL ダイバージェンスに基づく NMF（KL-divergence-based NMF: KLNMF）の生成モデルを明らかにしている [24]．KLNMF の最小化問題は，ポアソン生成モデルを仮定した最尤（maximum likelihood: ML）推定と等価である．式 (17) は，次の生成モデルを仮定している．

$$z_{imnj} \sim \mathcal{P}(z_{imnj}; a_{imn}s_{inj}) \tag{20}$$
$$\mathcal{P}(z;\lambda) = \frac{1}{\Gamma(z+1)} e^{-\lambda}\lambda^z \tag{21}$$

ここで，$z_{imnj} \in \mathbb{N}$ は $x_{imj} = e + \sum_n z_{imnj}$ を満たす確率変数，$\mathcal{P}(z;\lambda)$ は確率変数 $z \in \mathbb{N}$ とパラメータ $\lambda > 0$ を持つポアソン分布，$\Gamma(z+1) = z!$ はガンマ関数，e は $[0,1)$ の範囲で一様分布に従う確率変数である．また，z_{imnj} は i, m, n, j に関して互いに独立と仮定する．ポアソン分布に従う確率変数は再生性を持つ．即ち，$z_n \sim \mathcal{P}(z_n;\lambda_n)$ かつ $x = \sum_n z_n$ のとき，周辺分布は $p(x) = \mathcal{P}(x;\sum_n \lambda_n)$ で与えられる．従って，$\boldsymbol{X}_i$ の周辺対数尤度は次式となる．

$$\begin{aligned}
&\log p(\boldsymbol{X}_i;\boldsymbol{A}_i,\boldsymbol{S}_i) \\
&\quad = \log \prod_{m,j} \sum_{z_{imnj}} p(x_{inm};z_{imnj}) p(z_{imnj};a_{imn}s_{inj}) \\
&\quad = \log \prod_{m,j} \mathcal{P}(x_{imj};\textstyle\sum_n a_{imn}s_{inj}) \\
&\quad = \sum_{m,j} \Bigg[x_{imj} \log \sum_n a_{imn}s_{inj} \\
&\qquad - \sum_n a_{imn}s_{inj} - \log\Gamma(x_{imj}+1) \Bigg]
\end{aligned} \tag{22}$$

a_{imn} と s_{inj} に関する式 (22) の最大化（即ち最尤推定）は，式 (18) の最小化と等価な問題である [24]．

3.3 被り音の相対漏れゲインの事前分布生成モデル

提案手法では，$\boldsymbol{A}_i$ の自明解を避けるために，$\boldsymbol{A}_i$ の対角成分及び非対角成分に次の事前分布生成モデルを導入する．

$$a_{imn} \sim \begin{cases} \delta(a_{imn}-1) & (m=n) \\ \mathcal{G}(a_{imn};k,\theta) & (m \neq n) \end{cases} \tag{23}$$
$$\mathcal{G}(a;k,\theta) = \frac{1}{\Gamma(k)\theta^k} a^{k-1} e^{-a/\theta} \tag{24}$$

ここで，$\delta(a)$ は Dirac のデルタ関数であり，$\mathcal{G}(a;k,\theta)$ は確率変数 $a \geq 0$，形状母数 $k > 0$，及び尺度母数 $\theta > 0$ から成るガンマ分布である．また，a_{imn} は i，m，n に関して互いに独立と仮定する．式 (23) より，$\boldsymbol{A}_i$ の事前分布は次式となる．

$$\begin{aligned}
p(\boldsymbol{A}_i;k,\theta) &= \prod_{m,n=m} p(a_{imn}) \prod_{m,n\neq m} p(a_{imn};k,\theta) \\
&= \prod_{m,n=m} \delta(a_{imn}-1) \prod_{m,n\neq m} \mathcal{G}(a_{imn};k,\theta)
\end{aligned} \tag{25}$$

式 (25) では，$\boldsymbol{A}_i$ の対角成分は全て 1 となるよう制限され，非対角成分（被り音の相対漏れゲイン）の生起確率は k や θ によって制御される．Fig. 4 に示すように形状母数を $k > 1$ とすることで，$\boldsymbol{A}_i$ の全非対角成分において $a_{imn} = 0$ を避けることができ，$\boldsymbol{A}_i = \boldsymbol{I}$ なる自明解を回避できる．本稿では，以下 $k > 1$ のみを考える．

一方，アクティベーション行列 $\boldsymbol{S}_i$ には明示的な構造を仮定せず，次式の非負制約事前分布のみを導入する．

$$s_{inj} \sim \lim_{\beta\to\infty} \frac{1}{\beta}\mathcal{I}[0 \le s_{inj} \le \beta] \propto \mathcal{I}[0 \le s_{inj}] \tag{26}$$

ここで，β は正規化係数，$\mathcal{I}[\cdot]$ は引数が真ならば 1，偽なら 0 となる 2 値の関数である．$\boldsymbol{A}_i$ と同様に，s_{inj} も i，n，j に関して互いに独立と仮定されるため，$\boldsymbol{S}_i$ の事前分布は次式となる．

$$p(\boldsymbol{S}_i) = \prod_{n,j} p(s_{inj}) \propto \prod_{n,j} \mathcal{I}[0 \le s_{inj}] \tag{27}$$

3.4 MAP 推定のコスト関数

前節の事前分布に基づき，変数 $\boldsymbol{A}_i$ と $\boldsymbol{S}_i$ を MAP 推定で求める．事後分布は次のように得られる．

$$\prod_i p(\boldsymbol{A}_i, \boldsymbol{S}_i; \boldsymbol{X}_i) \propto \prod_i \underbrace{p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i)}_{\text{Likelihood}} \underbrace{p(\boldsymbol{A}_i; k, \theta) p(\boldsymbol{S}_i)}_{\text{Priors}} \tag{28}$$

式 (28) の負の対数を取ると，右辺を以下のように分解できる．

$$\mathcal{J} = -\sum_i [\log p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i) + \log p(\boldsymbol{A}_i; k, \theta) + \log p(\boldsymbol{S}_i)] \tag{29}$$

式 (22)，(25)，(27) を式 (29) に代入すると，コスト関数 $\mathcal{J}$ が次式として得られる．

$$\begin{aligned}\mathcal{J} = &\sum_{i,m,j} \left[-x_{imj} \log \sum_n a_{imn} s_{inj} \right. \\ &\left. + \sum_n a_{imn} s_{inj} + \log \Gamma(x_{imj} + 1) \right] \\ &+ \sum_{i,m,n=m} \mathbb{I}[a_{imn} = 1] \\ &+ \sum_{i,m,n\neq m} \left[-(k-1) \log a_{imn} + \frac{1}{\theta} a_{imn} \right] \\ &+ \sum_{i,n,j} \mathbb{I}[0 \le s_{imn}]\end{aligned} \tag{30}$$

ここで，$\mathbb{I}[\cdot] = -\log \mathcal{I}[\cdot]$ は引数が真のとき 0，偽のとき ∞ を取る指示関数である．$\boldsymbol{A}_i$ と $\boldsymbol{S}_i$ の MAP 推定は式 (30) の最小化問題であり，$\boldsymbol{A}_i$ と $\boldsymbol{S}_i$ に関する最小化は次の問題と等価である．

$$\begin{aligned}&\min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) + \sum_{i,m,n\neq m} \mathcal{R}(a_{imn}; k, \theta) \\ &\text{s.t. } a_{imn}, s_{inj} \ge 0 \ \forall i, m, n, j \text{ and } a_{imn} = 1 \ \forall m = n\end{aligned} \tag{31}$$

ここで，$\mathcal{R}(a_{imn}; k, \theta)$ は $\boldsymbol{A}_i$ の非対角成分のガンマ事前分布 (24) に対応する正則化項であり，次式となる．

$$\mathcal{R}(a_{imn}; k, \theta) = \left[-(k-1) \log a_{imn} + \frac{1}{\theta} a_{imn} \right] \tag{32}$$

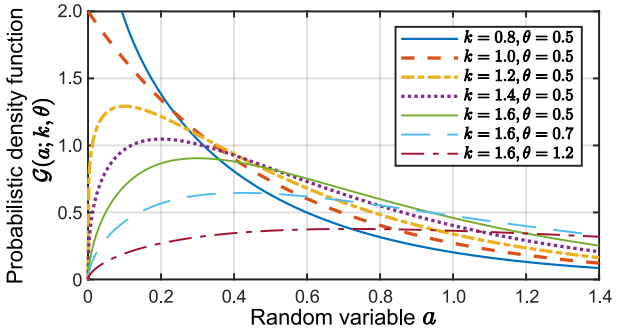

Fig. 4 Probabilistic density function of gamma distribution.

3.5 最適化アルゴリズムの導出

式 (31) の最小化問題は，NMF の最適化で頻繁に用いられる majorization-minimization（MM）アルゴリズム [18], [26] で解くことができる．目的関数 $\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i)$ の上限関数は Jensen's の不等式を用いて次のように設計される．

$$\begin{aligned}&\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) \\ &\overset{\mathrm{c}}{=} \sum_{i,m,j} \left(-x_{imj} \log \sum_n a_{imn} s_{inj} + \sum_n a_{imn} s_{inj} \right) \\ &= \sum_{i,m,j} \left(-x_{imj} \log \sum_n \xi_{imnj} \frac{a_{imn} s_{inj}}{\xi_{imnj}} + \sum_n a_{imn} s_{inj} \right) \\ &\le \sum_{i,m,j} \left(-x_{imj} \sum_n \xi_{imnj} \log \frac{a_{imn} s_{inj}}{\xi_{imnj}} + \sum_n a_{imn} s_{inj} \right) \\ &\equiv \mathcal{D}^+(\boldsymbol{A}_i, \boldsymbol{S}_i, \Xi)\end{aligned} \tag{33}$$

ここで，$\overset{\mathrm{c}}{=}$ は定数項の違いを除いて等しいことを表す．また，$\xi_{imnj} > 0$ は $\sum_n \xi_{imnj} = 1$ を満たす補助変数，Ξ は全ての i，m，j，n に対する ξ_{imnj} の集合を示す．式 (33) の等号成立条件は次式となる．

$$\xi_{imnj} = \frac{a_{imn} s_{inj}}{\sum_{n'} a_{imn'} s_{in'j}} \ \forall i, m, j, n \tag{34}$$

式 (33) より，上限関数の最小化問題は次のようになる．

$$\begin{aligned}&\min_{\mathcal{A},\mathcal{S},\Xi} \sum_i \mathcal{D}^+(\boldsymbol{A}_i, \boldsymbol{S}_i, \Xi) + \sum_{i,m,n\neq m} \mathcal{R}(a_{imn}; k, \theta) \\ &\text{s.t. } a_{imn}, s_{inj} \ge 0 \ \forall i, m, n, j, \quad \xi_{imnj} > 0 \ \forall i, m, n, j, \\ &\qquad \text{and } a_{imn} = 1 \ \forall m = n\end{aligned} \tag{35}$$

従って，式 (35) の上限関数を a_{imn} 及び s_{inj} で偏微分し 0 と置き，さらに式 (34) の統合成立条件を ξ_{imnj} に代入することで，各変数の更新式を導出できる．正則化項 $\mathcal{R}(a_{imn}; k, \theta)$ は s_{inj} での偏微分に影響しないため，s_{inj} の更新式は次式のように単純な KLNMF [18] の更新式と等しくなる．

$$s_{inj} \leftarrow s_{inj} \frac{\sum_m \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{inj'}} a_{imn}}{\sum_m a_{imn}} \tag{36}$$

一方，非対角成分 a_{imn}（$m \neq n$）については，式 (35) の上限関数の偏微分から次式が得られる．

$$\sum_j \left(-x_{imj} \frac{\xi_{imnj}}{a_{imn}} + s_{inj} \right) - (k-1) \frac{1}{a_{imn}} + \frac{1}{\theta} = 0 \tag{37}$$

式 (37) を a_{imn} について整理すると，次式となる．

$$a_{imn} = \frac{(k-1) + \sum_j x_{imj} \xi_{imnj}}{\frac{1}{\theta} + \sum_j s_{inj}} \tag{38}$$

式 (38) に統合成立条件 (34) を代入することで，非対角成分 a_{imn}（$m \neq n$）の更新式は次式として導出される．

$$a_{imn} \leftarrow \frac{(k-1) + a_{imn} \sum_j \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} s_{inj}}{\frac{1}{\theta} + \sum_j s_{inj}} \tag{39}$$

式 (36) 及び (39) より，a_{imn} 及び s_{inj} の非負制約は全変数の初期値を非負値とすることで常に満たされる．また，$\boldsymbol{A}_i$ の対角成分の値は制約条件を満たすように 1 で初期化し，他の変数の更新中も常に 1 に固定する．

式 (36) 及び (39) は，以下のように行列形式で実装すると，多くの言語で効率的に計算できる．

$$\boldsymbol{A}_i \leftarrow \frac{(k-1) + \boldsymbol{A}_i \odot \left(\frac{\boldsymbol{X}_i}{\boldsymbol{A}_i \boldsymbol{S}_i} \boldsymbol{S}_i^{\mathrm{T}} \right)}{\frac{1}{\theta} + \mathbf{1} \boldsymbol{S}_i^{\mathrm{T}}} \quad \forall i \tag{40}$$

$$\mathrm{diag}(\boldsymbol{A}_i) \leftarrow [1, 1, \cdots, 1]^{\mathrm{T}} \quad \forall i \tag{41}$$

$$\boldsymbol{S}_i \leftarrow \boldsymbol{S}_i \odot \frac{\boldsymbol{A}_i^{\mathrm{T}} \frac{\boldsymbol{X}_i}{\boldsymbol{A}_i \boldsymbol{S}_i}}{\boldsymbol{A}_i^{\mathrm{T}} \mathbf{1}} \quad \forall i \tag{42}$$

ここで，$\odot$ と行列の除算はそれぞれ要素毎の乗算と除算を表す．また，$\mathbf{1}$ は 1 だけを含む $M \times J$ の行列，$\mathrm{diag}(\cdot)$ は入力された正方行列の対角成分をベクトルとして表す．なお，式 (40) は $\boldsymbol{A}_i$ の対角成分の値を変更するが，式 (41) によって対角成分は直ちに 1 に上書きされる．また，式 (40) 及び (42) の反復計算は，コスト関数 (31) を単調減少させることが理論的に保証されている．

3.6 データ近似項と正則化項のバランス

提案手法では，$\boldsymbol{A}_i$ の対角成分を 1 に制約しているため，非対角要素は被り音の相対漏れゲインに対応する．また，KL ダイバージェンスは次式のスケール依存性を持つ．

$$\mathcal{D}_{\mathrm{KL}}(\alpha \boldsymbol{X}_i | \alpha \boldsymbol{A}_i \boldsymbol{S}_i) = \alpha \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) \tag{43}$$

ここで，$\alpha \geq 0$ は任意の係数である．これらの事実より，観測信号 $\{\boldsymbol{X}_i\}_{i=1}^I$ のゲインは，式 (31) におけるデータ近似項 $\sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i)$ 及び正則化項 $\sum_{i,m,n \neq m} \mathcal{R}(a_{imn}; k, \theta)$ 間のバランスに影響を与えることが分かる．そこで，提案手法では，観測信号を正規化したうえでゲインをパラメータ化する．具体的には，前処理として観測信号 $\tilde{\boldsymbol{x}}(t)$ に対して次のような処理を施す．

$$\tilde{\boldsymbol{x}}(t) \leftarrow \frac{\alpha}{v} \tilde{\boldsymbol{x}}(t) \quad \forall t \tag{44}$$

$$v = \max\left(\{\mathrm{abs}(\tilde{\boldsymbol{x}}(t))\}_{t=1}^T \right) \tag{45}$$

ここで，$\max(\cdot)$ は入力集合の最大値を返す．即ち，式 (44) は $\{\tilde{\boldsymbol{x}}(t)\}_{t=1}^T$ のダイナミックレンジが $\pm\alpha$ となるようにゲインを調整している．α は式 (19) における μ と同様に，データ近似項と正則化項のバランスを制御するハイパーパラメータとなる．例えば，α を小さな値に設定すると，最適化に対してより強い正則化が課せられる．

3.7 推定信号の再構成

提案手法では，従来の TCNMF と同様に，複素数の観測信号 x_{ijm} に対して次式の Wiener フィルタを適用することで，複素数の推定信号 $\mathbf{Y}_n$ を復元する．

$$\mathrm{y}_{ijn} = \frac{(a_{imm} s_{imj})^2}{\sum_n (a_{imn} s_{inj})^2} \mathrm{x}_{ijm} \tag{46}$$

なお，$a_{imm} = 1$ より，式 (46) は次式で実装できる．

$$\mathrm{y}_{ijn} = \left[\frac{\boldsymbol{S}_i^{:2}}{\boldsymbol{A}_i^{:2} \boldsymbol{S}_i^{:2}} \right]_{m,j} \mathrm{x}_{ijm} \tag{47}$$

ここで，$[\cdot]_{m,j}$ は入力行列の (m, j) 要素を表す．その後，推定信号 $\mathbf{Y}_n$ に逆 STFT を適用することで，時間領域の信号 $\tilde{y}_n(t)$ が得られる．推定信号のゲインは次式で復元する．

$$\tilde{\boldsymbol{y}}(t) \leftarrow \frac{v}{\alpha} \tilde{\boldsymbol{y}}(t) \quad \forall t \tag{48}$$

4. 実験

4.1 実験条件

提案手法の性能を評価するために，音楽信号の被り音抑圧の実験を行った．観測信号となる音楽の混合信号は，人工的な音楽データセットである *songKitamura* [27], [28] を用いた．混合前の音源信号 $\mathbf{S}_n$ には，クラリネット（Cl.），オーボエ（Ob.），ピアノ（Pf.），トロンボーン（Tb.）の 4 種類の楽器音を用い，$M = N = 4$ となるように 4 チャネルの観測信号 $\mathbf{x}_{ij}$ を作成した．このとき，被り音を含む観測信号を模擬するために，周波数毎の非負混合行列 $\overline{\boldsymbol{A}}_i \in \mathbb{R}_{\geq 0}^{M \times N}$ を用いて，楽器音 $\mathbf{s}_{ij}$ を次式のように混合した．

$$\mathbf{x}_{ij} = \overline{\boldsymbol{A}}_i \mathbf{s}_{ij} \tag{49}$$

但し，$\overline{\boldsymbol{A}}_i$ の対角要素は 1 とし，非対角要素は $(0, 0.2)$ の範囲の一様分布から生成される乱数に設定した．また，異なる擬似乱数シードを用いて，上記の方法で 10 種類の観測信号（即ち 10 種類の非負混合行列 $\overline{\boldsymbol{A}}_i$）を用いて実験を行った．信号のサンプリング周波数は 44.1 kHz とし，STFT では，窓長 4096 点（約 92.9 ms）の Hamming 窓をハーフ

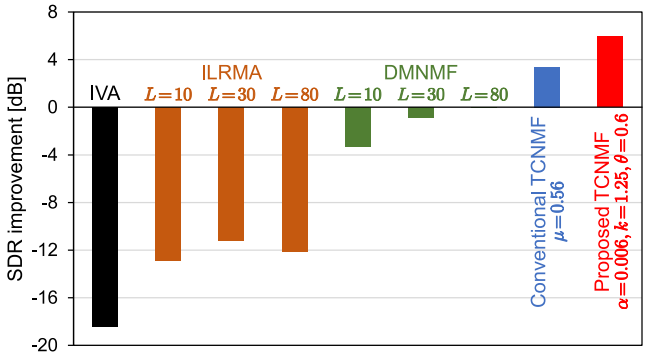

Fig. 5 Comparison of SDR improvements, where each bar is average over 10 different observed mixtures and 4 instrumental sources.

オーバーラップシフトさせた．周波数ビン数及び時間フレーム数はそれぞれ $I = 2049$ 及び $J = 109$ であった．

本実験では，独立ベクトル解析（independent vector analysis：IVA）[9]，ILRMA [11]，DMNMF [15]，従来の TCNMF [12]，及び提案手法の 5 つの手法を比較した．IVA 及び ILRMA は複素数の分離行列 $\mathbf{W}_i$ を推定する手法で位相を考慮した BSS であり，その他の手法は振幅又はパワースペクトログラムのみを用いる位相非依存な BSS である．IVA 及び ILRMA の $\mathbf{W}_i$ の初期値は，DMNMF や従来の TCNMF 及び提案手法で用いた初期混合行列の逆行列とした．また IVA と ILRMA では，反復音源ステアリング法（iterative source steering：ISS）[30] と呼ばれる数値的に安定な分離行列の更新式を用いた．分離行列 $\mathbf{W}_i$ 推定後は，式 (7) を用いて推定信号を復元し，さらにプロジェクションバック [31] を適用して周波数毎のスケールを復元した．一方，DMNMF，従来の TCNMF，及び提案手法については，非負混合行列 $\boldsymbol{A}_i$ の初期値に対角要素が 1，非対角要素が $(0, 0.1)$ の範囲の一様分布から生成される乱数を用いた．その他のパラメータについては，$(0, 1)$ の範囲で一様分布から生成される乱数を用いて初期化した．非負混合行列 $\boldsymbol{A}_i$ 推定後は，Wiener フィルタ (46) を用いて推定信号を得た．ILRMA と DMNMF における NMF 音源モデルの基底ベクトル数 L は 10，30，及び 80 と設定した．各手法の最適化アルゴリズムは更新式を 200 回反復計算し，いずれもコスト関数値が収束していることを確認している．

客観評価尺度として，source-to-interference ratio と sources-to-artificial ratio の 2 つを加味した総合的な分離品質を示す source-to-distortion ratio（SDR）[29] を用いた．1 章の条件 (a) で述べたように，被り音抑圧における観測信号の SDR（入力 SDR）は高い値を示した．今回の実験では，10 種類の観測信号の Cl.，Ob.，Pf.，及び Tb. の平均入力 SDR がそれぞれ 18.8，15.0，14.7，及び 8.6 dB となった．これらの入力 SDR からの改善量を楽器音毎に計算し，各手法での平均性能を評価した．

4.2 実験結果

Fig. 5 は 5 つの比較手法の平均 SDR 改善量を示している．ここで，Fig. 5 における従来の TCNMF と提案手法の性能は，ハイパーパラメータを実験的に決定し，本実験において最も性能が高かった $\mu = 0.56$，$k = 1.25$，$\theta = 0.6$，及び $\alpha = 0.006$ に設定した場合の結果を掲載している．位相を考慮した BSS である IVA と ILRMA では，被り音を全く抑圧できないことが確認できる．これは，本実験で観測された混合信号が，非負乱数混合行列 $\overline{\boldsymbol{A}}_i$ で生成されたものであることに起因しており，観測信号の位相が分離行列の推定に全く役に立たないためと考えられる．結果的に，IVA や ILRMA の推定信号には多くの人工的な歪みが生じ，SDR が大幅に劣化している．DMNMF は被り音を抑圧できる可能性があるが，その平均性能は 0 dB を超えなかった．これは，DMNMF における最適化の難しさに起因していると思われる．従来の TCNMF と提案手法では，平均 SDR 改善量が 0 dB を超えていることが確認できる．特に提案手法は，従来の TCNMF よりも 2.5 dB 以上改善されている．この改善は音楽演奏の高品質な後処理や SR を実現する上で重要といえる．

5. まとめ

本稿では，音楽の生演奏を近接マイクロホンで録音する際に生じる被り音の除去を目指し，TCNMF を改良した新しい手法を提案した．提案手法では，被り音の相対漏れゲインを正則化しており，これはガンマ事前分布を仮定した MAP 推定に基づく TCNMF と解釈できる．模擬的な観測信号を用いた実験では，提案手法が最も被り音を抑圧できることが実験的に示された．提案手法は 3 つのハイパーパラメータ（k，θ，及び α）を持っているため，ハイパーパラメータに対する提案手法の性能変化の解析と，効率的なパラメータのチューニング法が今後の課題として挙げられる．

謝辞

本研究の一部は，JSPS 科研費 19K20306 及び 19H01116 の助成を受けた．

参考文献

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag Berlin Heidelberg, 2001.
[2] H. L. Van Trees, *Optimum Array Processing*, John Wiley and Sons, New York, 2002.
[3] X. Yu, D. Hu, J. Xu, *Blind Source Separation: Theory and Applications*, John Wiley and Sons, New York, 2014.
[4] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, “A review of blind source separation methods: Two converging routes to ILRMA originating

from ICA and NMF," *APSIPA Trans. Signal and Info. Process.*, vol. 8, no. e12, pp. 1–14, 2019.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[6] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 2, pp. 666–678, 2006.

[7] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, pp. 601–608, 2006.

[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 189–192, 2011.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, S. Makino, Ed., pp. 125–155. Springer, Cham, 2018.

[12] M. Togami, Y. Kawaguch, H. Kokubo, and Y. Obuchi, "Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization," *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.*, pp. 522–525, 2010.

[13] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," *Proc. Int. Workshop Acoustic Signal Enhancement*, pp. 203–207, 2014.

[14] Y. Murase, H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "On microphone arrangement for multichannel speech enhancement based on nonnegative matrix factorization in time-channel domain," *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.*, 2014.

[15] T. Taniguchi and T. Masuda, "Linear demixed domain multichannel nonnegative matrix factorization for speech enhancement," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 476–480, 2017.

[16] O. Das, J. O. Smith, and J. S. Abel, "Microphone crosstalk cancellation in ensemble recordings with maximum likelihood estimation," *Proc. Audio Eng. Soc. Convention*, 2021.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization" *Proc. Neural Info. Process. Syst.*, pp. 556–562, 2000.

[19] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[20] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

[21] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, 2019.

[22] N. Makishima, Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Independent deeply learned matrix analysis with automatic selection of stable microphone-wise update and fast sourcewise update of demixing matrix," *Signal Process.*, vol. 178, 107753, 2021.

[23] T. Nakamura, S. Kozuka, and H. Saruwatari, "Time-domain audio source separation with neural networks based on multiresolution analysis," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 29, pp. 1687–1701, 2021.

[24] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. 785152, 2009.

[25] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[26] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, 2017.

[27] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 4, pp. 654–669, 2015.

[28] D. Kitamura, "Open dataset: songKitamura," `http://d-kitamura.net/dataset_en.html`. Accessed 17 August 2021.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[30] S. Robin and N. Ono, "Fast and stable blind source separation with rank-1 updates," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp.236–240, 2020.

[31] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.